International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Comparing Performances of Logistic Regression, Classification & Regression Trees and Artificial Neural Networks for Predicting Albuminuria in Type 2 Diabetes Mellitus

Imran Kurt Omurlu[a*], Mevlut Ture[b], Mustafa Unubol[c], Merve Katranci[d], Engin Guney[e]

[a,b,d]*Adnan Menderes University, Medical Faculty, Department of Biostatistics, Aydın, Turkey*

[c,e]*Adnan Menderes University, Medical Faculty, Division of Endocrinology, Aydın, Turkey*

[a]*E-mail: ikurtomurlu@gmail.com.*

Abstract

In this study, performances of classification methods were compared in order to predict the presence of albuminuria in type 2 diabetes mellitus patients. A retrospective analysis was performed in 266 subjects. We compared performances of logistic regression (LR), classification and regression trees (C&RT) and two artificial neural networks algorithms. Predictor variables were gender, urine creatinine, weight, blood urea, serum albumin, age, creatinine clearance, fasting plasma glucose, post-prandial plasma glucose, and HbA1c. For validation set, the best classification accuracy (84.85%), sensitivity (68.0%) and the highest Youden index (0.63) was found in the MLP model but the specificity was 95.12%. Additionally, the specificity of all the models was close to each other. For whole data set the results were found as 84.21%, 53.95%, 0.50 and 96.32% respectively. Consequently, the model had the highest predictive capability to predict the presence of albuminuria was MLP. According to this model, blood urea and serum albumin were the most important variables for predicting the albuminuria. On the basis of these considerations, we suggest that data should be better explored and processed

---

* Corresponding author.

E-mail address: ikurtomurlu@gmail.com.

by high performance modeling methods. Researchers should avoid assessment of data by using only one method in future studies focusing on albuminuria in type 2 diabetes mellitus patients or any other clinical condition.

Keywords: Albuminuria; Artificial Neural Networks; Classification and Regression Trees; Logistic Regression

1. Introduction

There are several risk factors associated with the type 2 diabetes mellitus. Among these overweight and obesity, sedentary lifestyle, previously identified glucose intolerance, age, gender, hypertension, history of gestational diabetes, decreased high-density lipoprotein, cholesterol, increased triglycerides, polycystic ovary syndrome, dietary factors, intrauterine environment, inflammation, family history of type 2 diabetes are the major components playing a critical role in the development of type 2 diabetes mellitus [1]. Common characteristics of a group of people are also at risk. This includes people of South Asian, African-Caribbean, black African and Chinese descent and those from lower socioeconomic groups [2]. Besides these modifiable and non-modifiable risk factors, urinary albumin is strong evident for diagnosis type 2 diabetes mellitus. Especially the higher albuminuria level is the primary predictor for development of diabetic nephropathy and for cardiovascular morbidity and mortality in patients with type II diabetes [3, 4, 5]. Current studies use classical statistical methods in order to predict the presence of albuminuria in type 2 diabetes mellitus patients. To predict with higher accuracy, the data should be explored and processed by high performance modeling methods.

Classic statistical classification methods have been usually used in classification problems when dependent variable is dichotomous. Researchers use data mining applications with higher accuracy and efficiency anymore, with popular classification methods like artificial neural networks (ANN), decision trees (DT) and random forests (RF) used for medical prediction [6]. These classification methods not only predict the outcome of a disease but also determine the predictor associated with outcome, relationships hidden deep into datasets and also specify the risk groups. Endo et al. [7] present optimal models to predict the survival of breast cancer patients. For this purpose they used ANN, logistic regression (LR), DT and Bayesian model by comparing their performances. Maroco et al. [6] compared discriminant analysis, LR, RF, classification trees, support vector machines, ANN (multilayer perceptron (MLP) and radial basis function (RBF)) for prediction of dementia patients. Ture et al. [8] compared various classification methods to predict control and hypertension groups. They created models using LR, flexible discriminant analysis, multivariate adaptive regression splines, chi-squared automatic interaction detector, quick unbiased efficient statistical tree, C&RT, RBF and MLP to predict hypertension. Morteza et al. [9] predicted albuminuria in patients with type 2 diabetes mellitus by using 2 different statistical models, MLP and conditional LR. Meng et al. [10] compared the performance of LR, ANN and DT models for predicting diabetes or prediabetes using common risk factors. Ture et al. [11] investigated the effect of some hypothetical factors on academic achievement using LR and chi-squared automatic interaction detector method. In study of Kurt et al. [12] the performances of classification methods (MLP, C&RT, LR, RBF and self-organizing feature maps) were compared in order to predict the presence of coronary artery disease.

The objective of this study is to compare performances of various classification methods to diagnose the presence of albuminuria in patients with type 2 diabetes mellitus. We created models using LR, C&RT, MLP

and RBF to predict albuminuria. We evaluated the classification accuracy, sensitivity, specificity and Youden index of these classification methods in another independent set of data records.

2. Material and Methods

2.1. Logistic regression

LR predicts the best model to describe the relationship between dependent variable and independent variables, like linear regression does. The difference between them, LR is used to predict a binary or dichotomous dependent variable [13].

The LR model for p independent variable is:

$$P = \frac{exp\,(\boldsymbol{\beta}'\boldsymbol{x})}{1+exp\,(\boldsymbol{\beta}'\boldsymbol{x})} \qquad (1)$$

where P is probability that dependent variable is in a particular category and $\boldsymbol{x} = (x_1, x_2, \dots, x_p)$ is a p-dimensional vector of independent variables and $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ are regression coefficients matrix. In this model, in order the estimate the parameters the maximum-likelihood method is used. There is a linear model hidden within the logistic regression model. The natural logarithm of the ratio of P to $(1-P)$ gives a linear model in x_i:

$$g(x) = ln\left(\frac{P}{1-P}\right) = \boldsymbol{\beta}'\boldsymbol{x} \qquad (2)$$

The$g(x)$, has many of the desirable properties of a linear regression model. The independent variables can be a combination of continuous and categorical variables [13].

2.2. Classification and regression trees

C&RT is a recursive partitioning method to be used both for regression and classification. C&RT is constructed by splitting subsets of the data set using all predictor variables to create two child nodes repeatedly, beginning with the entire data set. The best predictor is chosen using a variety of impurity or diversity measures (Gini, twoing, ordered twoing and least-squared deviation). The goal is to produce subsets of the data which are as homogeneous as possible with respect to the target variable [14]. In this study, we used measure of Gini impurity that used for categorical dependent variables.

Gini index at node t is defined as

$$Gini(t) = \sum_{j \neq i} p(j|t)p(i|t) \qquad (3)$$

where *i* and *j* are categories of the dependent variable. The equation for the Gini index can also be written as

$$Gini(t) = 1 - \sum_j p^2(j|t) \qquad (4)$$

Thus, when the cases in a node are evenly distributed across the categories, the Gini index takes its maximum value of $1-(1/k)$, where k is the number of categories for the dependent variable. When all cases in the node belong to the same category, the Gini index equals 0.

If costs of misclassification are specified, the Gini index is computed as

$$Gini(t) = \sum_{j \neq i} C(i|j)p(j|t)p(i|t) \quad (5)$$

where $C(i|j)$ is the probability of misclassifying a category j case as category i. The Gini criterion function for split s at node t is defined as

$$\Phi(s,t) = Gini(t) - p_L Gini(t_L) - p_R Gini(t_R) \quad (6)$$

where p_L is the proportion of cases in t sent to the left child node, and p_R is the proportion sent to the right child node. The split s is chosen to maximize the value of $\Phi(s,t)$. This value is reported as the improvement in the tree [14].

2.3. Multi-layer perceptron

MLP is a feed-forward network that the input propagates through the network via neurons from one layer to another in a forward direction [15, 16]. With one or more layers of hidden neurons the networks approximate the performance of optimal statistical classifiers in difficult problems [15].

The goal of MLP is to minimize the prediction error of one or more outputs. For this purpose, MLP trains the network by using a supervised learning technique called error back-propagation algorithm [15, 17].

For working with back propagation algorithm, an instantaneous error is defined. From the system response at neuron j at iteration t, $y_j(t)$, and the desired response $d_j(t)$ for a given input pattern the instantaneous error is calculate as

$$e_j(t) = d_j(t) - y_j(t) \quad (7)$$

The back propagation algorithm deals with second-order derivatives of the error surface are calculated by applying the Quasi-Newton algorithm is one of back propagation methods. Quasi-Newton algorithm is iterative method that involves a series of line searches. The method at the each iteration computes the change in weight and takes the search direction at t^{th} iteration [18]. The change in weight:

$$w_{jk}(t+1) = w_{jk}(t) - \eta(t)s(t) \quad (8)$$

The $\eta(t)$ is the learning-rate parameter. The $w_{jk}(t)$ is the weight connecting the output of neuron k to the input neuron j at iteration t. The $s(t)$ is direction vector and calculated as $s(t) = -S(t)g(t)$ by using positive definite matrix $S(t)$ and gradient vector $g(t)$ [15].

2.4. Radial basis function

The structure of RBF networks bases on two layer feed-forward network. In between input and output layer there is single hidden layer which applies a nonlinear transformation from the input space to hidden space by using radial basis activation function, in contrast to classical neural networks. This nonlinear transformation follows linear transformation of output layer [15, 19, 20]. The advantage of the radial basis function network is that it finds the input to output map using local approximations. The key for a successful implementation of these networks is to find suitable centers for the Gaussian functions [15, 21].

In the RBF, each node in the hidden layer is a p multivariate Gaussian function

$$G(x; C_r) = e^{-\left(\frac{\|x-C_r\|}{2\sigma_r^2}\right)^2} \quad \text{(r=1,2,...,k)} \qquad (9)$$

of mean C_r (each data point) and σ_r variance. These functions are called radial basis functions. The k is the number of neurons (cluster centers) at hidden layer. Finally, linearly weight the output of the hidden neuron to obtain:

$$F(x) = \sum_{r=1}^{k} w_r G(x; C_r) \qquad (10)$$

The problem with this solution is that it may lead to a very large hidden layer [15, 21].

2.5. Albuminuria data

Urinary albumin loss can be categorized into 3 classes depending on the amount of albumin lost. Nonalbuminuria is defined as a urinary albumin loss of 0-30 mg/24 hours, microalbuminuria as 30-300 mg/24 hours and macroalbuminuria as ≥300 mg/24 hours [22]. Diabetic nephropathy develops in 20-40% of diabetic individuals [4]. Microalbuminuria, a reversible phase of diabetic nephropathy, is an important finding and characterized by a 30-299 mg/L excretion of albumin in 24-hour urine [23]. Microalbuminuria is the primary predictor for development of diabetic nephropathy in patients with type II diabetes [4, 5]. We used to analyze the albuminuria data in type 2 diabetes mellitus from Guney et al. [24]. In total there were 10 predictors and 266 cases in the data. The dependent variable was a binary categorical variable with two categories: nonalbuminuria and micro+macro albuminuria. The distribution of the dependent variable is shown in the Table 1. Table 2 shows the summaries of predictor variables. In all cases, gender, urine creatinine (mg/dl), weight (kg), blood urea (mg/dl), serum albumin (mg/dl), age (year), creatinine clearance (mL/min), fasting plasma glucose (mg/dl), post-prandial plasma glucose (mg/dl), and HbA1c (mg/dl) were assessed and documented.

Before analyzing models, to find the optimum models, we divided the data set into 3 different sets (training, test and validation sets). The training set was used to build a model and to discover a predictive relationship for prediction the presence of albuminuria in type 2 diabetes mellitus patients. In the ANN, learning process is continued iteratively, and the network progresses to improve its predictions until the stopping criteria have been met. Thus, for MLP and RBF, the test set was used as early stopping criterion of network training in order to

avoid over fitting. The validation set as an independent set of data records was used to assess performance of four building models. Therefore, the data set were randomly split into three subsets, 60% of the data for training set, 20 % of data for test set and 20% of data for validation set (Fig. 1).

Table 1. Distribution of dependent variable

Category	Frequency	%
Nonalbuminuria	190	71.4
Micro+macro albuminuria	76	28.6
Total	266	100

Table 2. Descriptive statistics of predictor variables

Categorical variable name	Frequency (%)
Gender (M/F)	121/145 (45.5/54.5)
Continuous variable name	Mean ± standard deviation or Median (25-75 percentiles)
Age (year)	56.16±10.22
Weight	79.29±14.49
Serum albumin	4.41±0.49
Fasting plasma glucose	139.00 (110.00-194.50)
Blood urea	28.00 (23.00-35.00)
HbA1c	7.50 (6.40-9.30)
Creatinine clearance	119.50 (89.75-144.25)
Urine creatinine	1345.71±503.69
Post-prandial plasma glucose	192.00 (136.75-281.00)

The MLP was employed for structuring model and configured with one input layer, one hidden layer and one output layer. This method was performed under varying the number of hidden layer units (range between 2-200). The best number of hidden units from the range was determined using the testing data criterion. To control the training, the momentum term was set at 0.9 used to speed up and stabilize convergence to the weight update. The learning rate was set at 0.3 as the starting value so as to control the changed weights. It decreases to 0.01 as low eta and reset to 0.1 as high eta and decreases to low eta again. This repeating process is known as eta decay is the number of cycles and in this study eta decay was set at 30. The RBF was performed under varying the number of centers (range between 20-30). The momentum term was set at 0.9 as in MLP model. But it remains constant

during the process in contrast to MLP. The learning rate was computed automatically by the model based on the number of cycles in the repeating process. To control how much clusters overlapped, the RBF overlapping was set at 1. In decision tree construction by the C&RT model we used Gini impurity measure in order to choose the best predictor for albuminuria data. In producing the LR equation, the maximum-likelihood ratio was used to determine the statistical significance of the variables. Fit was assessed by Hosmer-Lemeshow test as goodness-of-fit test.

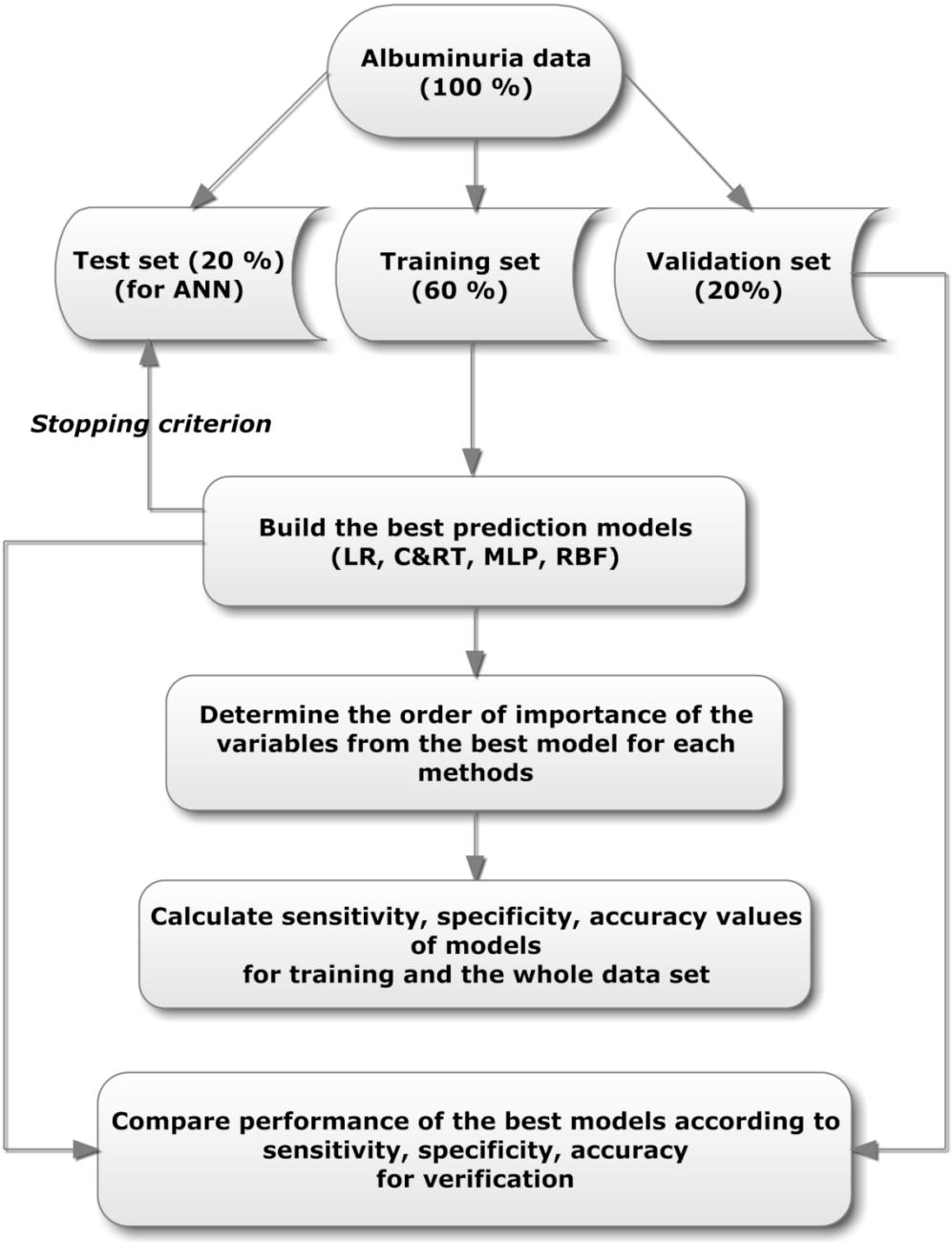

Fig. 1. a flowchart representation for classification methods

To determine the order of importance of each variable in four models was calculated the importance value. The importance value close to zero represents that a poor relationship between the particular variable and correct classification. To find the optimal model for predict the presence of albuminuria in type 2 diabetes mellitus we appraised the classification methods using confusion matrix for accuracy, sensitivity, and specificity and Youden index. Sensitivity is the ratio of people with disease correctly identified (actual patients) and specificity refers to the ratio of people without disease correctly identified whereas accuracy measures how well the model correctly identifies both people with disease and without disease:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

where TP, TN, FP, and FN represents true positive, true negative, false positive and false negative, respectively.

Youden's index is a measure for classification accuracy of diagnostic test and calculated by sensitivity and specificity values of the test:

Youden index = Sensitivity + Specificity - 1

The index takes a value between -1 and 1. If the value is smaller than 0, the test indicates that it isn't sufficient for diagnosis. If Youden indices of two tests are compared, value the bigger index has the highest accuracy.

3. Results

Using the independent predictors, we used LR analysis with p=0.05 for entry in forward stepwise variable selection to predict the probability of presence of albuminuria. Fit was assessed by chi-square statistics proposed by Hosmer-Lemeshow goodness-of-fit test, and LR model showed a good fit statistics (χ^2=10.41, p>0.05). In Table 3, we gave estimates, Wald test statistics and odds ratio (OR) of the regression coefficients in the forward stepwise LR. We found that gender (p=0.002, OR=0.405), blood urea (p=0.005, OR=1.030), and serum albumin (p=0.001, OR=0.321) had significant effects on albuminuria.

Table 3. The independent variables in LR model

Variable	Coefficient	Wald	p	OR	95 % C.I. for OR	
					Lower	Upper
Gender / Male	-0.904	9.250	0.002	0.405	0.226	0.725
Blood urea	0.030	7.852	0.005	1.030	1.009	1.052
Serum albumin	-1.136	10.580	0.001	0.321	0.162	0.637
Constant	3.531	4.737	0.030	34.170		

The decision rules of C&RT provide specific information about risk factors based on the rule induction. The C&RT has 12 leaf nodes, of which 7 are terminal nodes. The variables were: blood urea, post-prandial plasma glucose, fasting plasma glucose, urine creatinine and serum albumin. The blood urea was the most important determining factor for albuminuria. This first-level split produced the two initial branches of the decision tree. As shown in Table 4, for the blood urea (≤40.5) subgroup, post-prandial plasma glucose status proved the best predicting variable; post-prandial plasma glucose (>495) versus post-prandial plasma glucose (≤495). For post-prandial plasma glucose (>495) subgroup there was no predicting variable so it was the terminal node. For the post-prandial plasma glucose (≤495) subgroup urine creatinine status proved the best predicting variable, urine creatinine (>1247) versus urine creatinine (≤1247). For the urine creatinine (>1247) subgroup fasting plasma glucose status proved the best predicting variable, fasting plasma glucose (≤94) versus fasting plasma glucose (>94). For fasting plasma glucose subgroup there was no predicting variable. For the blood urea (>40.5) subgroup serum albumin status proved the best predicting variable; serum albumin (>4.45) versus serum albumin (≤4.45). For serum albumin (≤4.45) subgroup there was no predicting variable so it was the terminal node. For the serum albumin (>4.45) subgroup fasting plasma glucose status proved the best predicting variable, fasting plasma glucose (>192.5) versus fasting plasma glucose (≤192.5). For serum albumin subgroup there was no predicting variable.

In the MLP methods, the number of the optimum hidden units was found 2 units. For prediction of albuminuria, the optimum RBF network that consisted of 20 centers in its hidden layer was found (Table 5).

Table 4. Terminal nodes of decision tree for C&RT

Node	Terminal node
4	blood urea (≤40.5)+ post-prandial plasma glucose (>495)
5	blood urea (>40.5) + serum albumin (≤4.45)
7	blood urea (≤40.5) + post-prandial plasma glucose (≤495) + urine creatinine (≤1247)
9	blood urea (>40.5) + serum albumin (>4.45) + fasting plasma glucose (≤192.5)
10	blood urea (>40.5) + serum albumin (>4.45) + fasting plasma glucose (>192.5)
11	blood urea (≤40.5) + post-prandial plasma glucose (≤495) + urine creatinine (>1247) + fasting plasma glucose (≤94)
12	blood urea (≤40.5) + post-prandial plasma glucose (≤495) + urine creatinine (>1247) + fasting plasma glucose (>94)

Table 5. Summary of MLP and RBF Networks

Networks	The number of hidden units	Classification Rate	
		Training	Test
MLP	2	84.77	81.63
RBF	20	78.15	71.43

To detect placed in order of importance of each variable, the importance of the all input variables performed for models was showed in Table 6. As shown in the table, albumin was the most important variable in both LR and RBF models. The variable had the highest importance in MLP was blood urea and it followed by serum albumin. Blood urea was also the most important variable in C&RT while it had lower importance in RBF model. Urine creatinine was the second important variable in both C&RT and RBF model. While gender played an important role in LR model after serum albumin, it had lower importance in other models. Generally, the most important variables were found as blood urea and serum albumin in albuminuria dataset.

Table 6. The order of all independent variables according to importance

Order	LR	C&RT	MLP	RBF
1	Serum albumin	Blood urea	Blood urea	Serum albumin
2	Gender	Urine creatinine	Serum albumin	Urine creatinine
3	Blood urea	Post-prandial plasma glucose	Age	Fasting plasma glucose
4	Age	Serum albumin	Urine creatinine	Weight
5	Weight	HbA1c	Fasting plasma glucose	HbA1c
6	HbA1c	Gender	Creatinine clearance	Gender
7	Creatinine clearance	Weight	Post-prandial plasma glucose	Age
8	Fasting plasma glucose	Creatinine clearance	Weight	Blood urea
9	Urine creatinine	Fasting plasma glucose	Gender	Post-prandial plasma glucose
10	Post-prandial plasma glucose	Age	HbA1c	Creatinine clearance

A comparison of the sensitivity, specificity, accuracy and Youden index for albuminuria data set of classification methods were shown in Table 7. All models had sensitivity, specificity, accuracy and Youden range between 16.67-50.00%, 97.39-99.13%, 78.15-87.42% and 0.14-0.49 respectively for training set. For validation set the models had sensitivity, specificity, accuracy and Youden range between 32.0-68.0%, 92.68-97.56%, 69.70-84.85% and 0.25-0.63. In whole dataset sensitivity, specificity, accuracy and Youden range between 26.32-

53.95%, 96.32-97.89%, 76.31-84.21% and 0.23-0.50 respectively. The specificity was more than 90% in each model and set. In this study, C&RT model had the highest sensitivity (50.00%), specificity (99.13%), the best classification accuracy (87.42%) and the highest Youden index (0.49) for training set. For validation set the best classification accuracy (84.85%) was found in the MLP model with the highest Youden index. MLP model had also the best classification accuracy (84.21%) and the highest Youden index for whole dataset. The best model according to sensitivity (68.0%) was MLP while the RBF model gave the best specificity (97.56%) in validation set. For whole data set the model had the highest specificity (97.89%) was LR model while the MLP model had the highest sensitivity (53.95%). On the basis of the comparison results, the MLP model was found the best classification method among four methods when models were assessed for their verifying capacities by using validation set.

Table 7. Comparison of the performance according to classification matrix and predicted values of models

Set	Models	Sensitivity (%)	Specificity (%)	Accuracy (%)	Youden
Training	LR	22.22	99.13	80.79	0.21
	C&RT	50.00	99.13	87.42	0.49
	MLP	44.44	97.39	84.77	0.42
	RBF	16.67	97.39	78.15	0.14
Validation	LR	40.00	95.12	74.24	0.35
	C&RT	32.00	92.68	69.70	0.25
	MLP	68.00	95.12	84.85	0.63
	RBF	40.00	97.56	75.76	0.37
The whole dataset	LR	28.95	97.89	78.20	0.27
	C&RT	38.16	96.32	79.70	0.34
	MLP	53.95	96.32	84.21	0.50
	RBF	26.32	96.32	76.31	0.23

4. Conclusion

The goal of this study is to develop an accurate and comprehensible classifier for prediction albuminuria. Therefore, we used LR, C&RT and two different ANN algorithms (RBF and MLP) on this data, and compared them to each other using performance measurements. We aimed to discover the risk factors and to make decision rules for the management of albuminuria in type 2 diabetes mellitus patients by using these models. In all cases, gender, urine creatinine, weight, blood urea, serum albumin, age, creatinine clearance, fasting plasma glucose, post-prandial plasma glucose, and HbA1c were assessed. Considering the ordered importance of variables in Table 6, generally the most important variables were found as blood urea and serum albumin in albuminuria dataset. This provides us information about general tendency of data structure for prediction of albuminuria.

However, when RBF network was applied to the data, blood urea had minor role in the prediction of albuminuria than other variables compared other models.

The comparison of classification methods in different dataset has been presented in previous studies as follows. Delen et al. [25] compared support vector machines model with RBF and polynomial, MLP, RBF, M5 and C&RT as prediction models for prognostic analysis of thoracic transplantations and exploring risk groups. In order to identify the risk groups Cox regression analysis was applied and the discrimination among the risk groups was validated by Kaplan-Meier survival analysis. In consequence of analyzes, they reported that support vector machines model with RBF was the best model, MLP model came the second and C&RT model came the last. Kurt et al. [12] created models using LR, MLP, RBF, C&RT and self-organizing feature maps models to predict coronary artery disease by retrospective analysis in 1245 subjects. They used not only ROC curve, but also hierarchical cluster analysis and multidimensional scaling to compare performances of these models. Consequently they found that MLP was the best model with its good classificatory performance to predict presence of coronary artery disease. They suggested that not only area under the ROC curve was preferred in comparing performances of the models but also hierarchical cluster analysis and multidimensional scaling could be use since they provide us a visual assessment. Oludolapo et al. [26] found that the RBF model was a more accurate predictor than the MLP model for calculating the energy consumption of South Africa's industrial sector. García et al. [27] presented that a comparison among three different classifiers (LR, MLP, RBF) in the detection of hard exudates in retinal images. They obtained better results using MLP and RBF than using LR model. Ramana et al. [28] compared naïve bayes classifier, C 4.5, ANN, k-nearest neighbor and support vector machines for classifying liver patients in liver disease diagnosis. They also compared two dataset to compare the techniques. Consequently, k-nearest neighbor, ANN and support vector machines gave better results both two dataset. In other study of Ramana et al. [29] the modified rotation forest was used for liver classification. They compared it with tree based, statistical based, ANN based, rule based, lazy learners and support vector machines. According to comparison models MLP gave the highest accuracy for liver classification. Yasin et al. [30] proposed a new classification approach for hepatitis-c diagnosis. They compared it with nineteen classification techniques used in previous different studies. According to classification techniques obtained by using hepatitis diagnostic methods, their approach had the highest accuracy. Polat et al. [31] proposed a technique based on generalized discriminant analysis and least square support vector machine to diagnosis of diabetes disease. According to obtained classification accuracies, they reported that the combination of these techniques had the highest accuracies compared to the previously reported classification techniques.

Comparing performances of the models based on validation set is an appropriate assessment for prediction in this data, because according to the study of Simon et.al [32] the predictors in the original data may not accurately reflect all characteristics of the underlying populations of interest, therefore the data require verification of the predictors on independently created set of data. In our study, considering the results of validation set, the ANN models performed better than LR model and the LR model performed better than C&RT model. We concluded that our model based on MLP achieved the best performance, with 84.85% accuracy. The RBF model came out second, with 75.76% accuracy and it followed by LR model with 74.24% accuracy. The C&RT model was poorest in four models, with 69.70% accuracy.

Each technique shows some characteristics which may be interesting in the context of clinical practice. For instance, LR is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of set of predictor variables which are continuous, categorical, or both. Furthermore, it assumes that measures of dependent variables are independently and randomly sampled, all potentially relevant independent variables are in the model and all independent variables in the model are relevant [13, 33]. The tree representation in C&RT is close to the medical reasoning and can help to structure the understanding of prediction. Although artificial neural networks are mathematical models they overcome wide variety of complex tasks that are hard to solve. These methods probably have the potential to complement existing statistical models and to contribute to the interpretation and presentation of data.

In our study, we compared methods by using a real data set in order to predict albuminuria and provide information about which of those predictors played major role in the data. Another of our scopes in this study is to help researchers to select best method for solving problems of classification especially in albuminuria data in type 2 diabetes mellitus also any other medical conditions.

References

[1] K.G.M.M. Alberti, P. Zimmet, J. Straw. "International diabetes federation: A consensus on type 2 diabetes prevention". *Diabet Med*, vol. 24, pp. 451-463, 2007.

[2] National Institute for Health and Clinical Excellence (NICE) public health guidance 35. *Preventing Type 2 Diabetes: Population and Community Level Interventions*, pp. 1-91, 2011.

[3] D. de Zeeuw. "Albuminuria: A target for treatment of type 2 diabetic nephropathy". *Semin Nephrol*, vol. 27, pp. 172-181, 2007.

[4] American Diabetes Association. "Standards of medical care in diabetes". *Diabetes Care*, vol. 33, pp. 11-61, 2010.

[5] M. Unubol, M. Ayhan, E. Guney. "The relationship between mean platelet volume with microalbuminuria and glycemic control in patients with type II diabetes mellitus". *Platelets*, vol. 23, pp. 475-80, 2012.

[6] J. Maroco, D. Silva, M. Guerreiro, A. de Mendonça, I. Santana. "Prediction of dementia patients: A comparative approach using parametric vs. non parametric classifiers," in *Proc.* XIX Congresso Anual da Sociedade Portuguesa de Estatistica, Portuguese, 2011.

[7] A. Endo, T. Shibata, H. Tanaka. "Comparison of seven algorithms to predict breast cancer survival". *Biomedical Soft Computing and Human Sciences*, vol. 13, pp. 11-16, 2008.

[8] M. Ture, I Kurt, A.T. Kurum, K. Ozdamar. "Comparing classification techniques for predicting essential hypertension". *Expert Syst Appl*, vol. 29, pp. 583-588, 2005.

[9] A. Morteza, M. Nakhjavani, F. Asgarani, F.L.F Carvalho, R. Karimi, A. Esteghamati. "Inconsistency in albuminuria predictors in type 2 diabetes: A comparison between neural network and conditional logistic regression". *Translational Research*, vol. 161, pp. 397-405, 2013.

[10] X. Meng, Y. Huang, D. Rao, Q. Zhang, Q. Liu. "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors". *Kaohsiung J Med Sci*, vol. 29, pp. 93-99, 2013.

[11] M. Ture, Z. Akturk, I. Kurt, N. Dagdeviren. "The effect of health status, nutrition, and some other factors on low school performance using induction technique". *Trakya Univ Tip Fak Derg*, vol. 23, pp. 28-38, 2006.

[12] I. Kurt, M. Ture, A.T. Kurum. "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease". *Expert Syst Appl*, vol. 34, pp. 366-374, 2008.

[13] D.W. Hosmer, S. Lemeshow. *Applied Logistic Regression*, New York: Wiley, 2000.

[14] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. *Classification and Regression Trees*. New York: Chapman and Hall/CRC, 1984, pp. 93-126.

[15] S. Haykin. *Neural Network: A Comprehensive Foundation*, Upper Saddle River, NJ: Prentice Hall, 1999.

[16] B. Kröse, P. van der Smagt. *An Introduction to Neural Networks*. University of Amsterdam, 1996.

[17] G. Calcagno, A. Staiano, G. Fortunato, V. Brescia-Morra, E. Salvatore, R. Liguori, et al. "A multilayer perceptron neural network based approach for the identification of responsiveness to interferon therapy in multiple sclerosis patients". *Inform Sciences*, vol. 180, pp. 4153-63, 2010.

[18] P. Hennig, M. Kiefel. "Quasi-Newton methods: A new direction". *J Mach Learn Res*, vol. 14, pp. 843-65, 2013.

[19] A.G. Bors. "Introduction of the radial basis function (RBF) networks," in *Proc.* Online Symposium for Electronics Engineers, vol. 1, pp. 1-7, 2001.

[20] M.H. Hassoun. *Fundamentals of Artificial Neural Networks*, MIT Press, Cambridge, 1995.

[21] J. Principe, N.R. Euliano, W.C Lefebvre. *Neural and adaptive systems: Fundamentals through simulations*. New York: Wiley, 1999.

[22] P.E. de Jong, R.T. Gansevoort, S.J. Bakker. "Macroalbuminuria and microalbuminuria: Do both predict renal and cardiovascular events with similar strength?" *J Nephrol*, vol. 20, pp. 375-80, 2007.

[23] D.E. Busby, G.L. Bakris. "Comparison of commonly used assays for the detection of microalbuminuria". *J Clin Hypertens*, vol. 6, pp. 8-12, 2004.

[24] E. Guney, M. Unubol, V. Yazak, I. Kurt Omurlu. "Tip 2 diyabetli hastalarda albüminürisiz nefropatiyi yakalayabiliyor muyuz?" in *Proc*. 47. Ulusal Diyabet Kongresi, Antalya, pp.123, 2011.

[25] D. Delen, A. Oztekin, Z.J. Kong. "A machine learning-based approach to prognostic analysis of thoracic transplantations". *Artif Intell Med*, vol. 49, 33-42, 2010.

[26] A.O. Oludolapo, A.A. Jimoh, P.A. Kholopane. Comparing performance of MLP and RBF neural network models for predicting South Africa's energy consumption". *Journal of Energy in Southern Africa*, vol. 23, pp. 40-6, 2012.

[27] M. García, C. Valverde, I.M. López, J. Poza, R. Hornero. "Comparison of logistic regression and neural network classifiers in the detection of hard exudates in retinal images," in *Proc.* 35th Annual International Conference of the IEEE EMBS, Osaka, Japan, pp. 3-7, 2013.

[28] B.V. Ramana, M.S.P. Babu, N.B. Venkateswarlu. "A critical study of selected classification algorithms for liver disease diagnosis". *International Journal of Database Management Systems*, vol. 3, pp. 101-114, 2011.

[29] B.V. Ramana, M.S.P. Babu. "Liver classification using modified rotation forest". *International Journal of Engineering Research and Development*, vol. 1, pp. 17-24, 2012.

[30] H. Yasin, A.T. Jilani, M. Danish. "Hepatitis-C classification using data mining techniques". *Int J Comput Appl*, vol.24, pp. 1-6, 2011.

[31] K. Polat, S. Gunes, A. Arslan. "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine". *Expert Syst Appl*, vol. 34, pp. 482-7, 2008.

[32] R. Simon, D.M. Radmacher, K. Dobbin, M.L. McShane. "Pitfalls in the Use of DNA microarray data for diagnostic and prognostic classification". *J Natl Cancer Inst*, vol. 95, pp. 14-8, 2003.

[33] D.G. Kleinbaum. *Logistic regression: A self-learning text.* Springer-Verlag, New York, 1994.